LABORATORY FOR
COMPUTER SCIENCE

MASSACHUSETTS
INSTITUTE OF
TECHNOLOGY

MIT/LCS/TR-495

THE SPECTRAL NORM OF FINITE FUNCTIONS

Mihir Bellare

February 1991

545 TECHNOLOGY SQUARE, CAMBRIDGE, MASSACHUSETTS 02139

This document for written for my area exam. The assigned papers were [FJS],[KM] and [SB].

The Spectral Norm of Finite Functions

Mihir Bellare

Laboratory for Computer Science
Massachusetts Institute of Technology

February 1, 1991

Abstract

In many recent results in learning and computational complexity theory which rely on Fourier analysis, the *spectral norm* plays a key role. An understanding of this quantity would appear to be useful in both gauging and exploiting these results, and in understanding the underlying techniques.

This paper surveys various aspects of the spectral norm of finite functions. We consider some of the motivating results as well as both upper and lower bounds on the spectral norm and their relationships to the computational complexity of the function.

Some of the results included here are new. In particular, we introduce a general technique for upper bounding the spectral norm of a decision tree over an arbitrary basis. We also extend the learning algorithm of Kushilevitz and Mansour [KM] to mutually independent distributions.

Keywords: Fourier series, spectral norm, learning, decision trees.

Contents

1 Introduction

Most functions are hard to learn efficiently. Functions of *small spectral norm*, say Kushilevitz and Mansour [KM], are an exception: they can be learned in polynomial time.

Moreover these functions can be approximated by sparse Fourier series [SB]. They are also computable by threshold circuits of very small depth [Br],[HMPST],[BS],[SB]. And so on: the list of their virtues is a long one.

Small is indeed beautiful. And that is just one reason to be interested in spectral norms.

FOURIER SERIES AND THE SPECTRAL NORM. The Fourier series of a function $f : \{0,1\}^n \to \mathbf{R}$ is obtained by expressing it as a linear combination of certain appropriately chosen *basis functions*: $f = \sum_{z \in \{0,1\}^n} \widehat{f}(z)\chi_z$ where the $\chi_z : \{0,1\}^n \to \{-1,+1\}$ are the basis functions and the $\widehat{f}(z)$ are real numbers depending on f. The spectral norm of f, denoted $L(f)$, is the sum of the absolute values of the coefficients in this linear combination: $L(f) = \sum_{z \in \{0,1\}^n} |\widehat{f}(z)|$.

Fourier series have of late proved to be a very useful tool in learning and computational complexity theory. And as the above indicates, the spectral norm often plays a pivotal role.

It is not just that small is beautiful. If one examines these results more closely, one realizes that the spectral norm is a kind of "yardstick," its value for a given function measuring the "complexity" of this function with respect to the task at hand. For example, what the [KM] learning result really says is that a function is learnable in time polynomial in its spectral norm. The sparse approximation result [SB] really says that one can approximate a function by a Fourier series having a number of terms proportional to the spectral norm of the function. And so on.

The value of the spectral norm of a function emerges as a key to many of its computational properties. In order to exploit results such as those listed above, it becomes important to know more about this value. In particular, which functions have small spectral norm? Which don't? How does this relate to their complexity in various models of computation? These are the kinds of questions on which I will focus.

THIS REPORT. I begin by describing a couple of the motivating results mentioned above.

Next I look at upper bounds on the spectral norm, where what one is really interested in, of course, is to identify the functions of small norm. Here I present two very simple examples.

I then look at the spectral norm of decision trees, an interesting case of how the spectral norm of a function can relate to its complexity in some model of computation.

In §6 I look at lower bounds, presenting two examples of functions whose spectral norm is close to the highest possible value for any boolean function.

Finally, I turn to a generalization of the usual spectral norm which I call the q-norm. Results about the q-norm being much scarcer in the literature (it is a recent invention of Furst, Jackson and Smith [FJS]) I will spend more time here on developing techniques and applications. I conclude by considering briefly the relative merits of the two norms.

WHAT'S NEW. Among the results and proofs included here, some are new. The principal amongst these are

- A technique for upper bounding the spectral norm of a decision tree over an *arbitrary* basis.

  I introduce a general framework which enables one to compute, in a simple manner, a *weight*

which for any decision tree is an upper bound on its spectral norm. This extends the result of [KM] which applied only to decision trees over the *parity* basis. As a consequence, many new classes of decision trees become learnable in polynomial time.

- An extension of the [KM] learning algorithm to mutually independent distributions.

  I generalize the proof of [KM] to show that their algorithm works to learn functions in time polynomial in their q-norm when the error of a hypothesis is measured by its proximity to the target under the mutually independent distribution q.

- A new proof of an exponential (and almost optimal) lower bound on the spectral norm of the majority function.

  This is only new in the sense that I have not seen any other proof (however I know that the result is known: I have just been unable to track down the reference).

APOLOGIA. It being the first time I have looked at Fourier analysis, I have not hesitated to state the obvious. In order to tie results together and set the stage, I have often had to prove small things which are not in the papers only, I'm sure, because they are considered too basic.

I include my proofs anyway. To disambiguate, I have followed the convention of citing in the theorem statement the source from which I took it: if nothing is cited it means I couldn't find it anywhere.

Finally, I must apologize for the many things I have omitted. It seemed better to stick to a single thread than to try to compile a compendium of results in the assigned papers, and of course much that is interesting is lost.

2 Basics

The set of real valued functions on $\{0,1\}^n$ is a 2^n dimensional vector space over the reals. Choosing the "right" basis for this vector space yields the Fourier series.

Here I'll go through the basics, as briefly as possible. I'll give the definitions and a minimum of the rich set of properties of Fourier series, developing more as the need arises in later sections.

NOTATION AND CONVENTIONS. I identify a string $z \in \{0,1\}^n$ with the subset of $\{0,1\}^n$ whose charecteristic vector is z. With this convention, I will apply set-theoretic notation to strings, using expressions like "$i \in z$" or "$y \triangle z$" where y, z are n bit strings and $1 \leq i \leq n$.

The empty string is denoted λ. We adopt the (usual) convention that $\{0,1\}^0 = \{\lambda\}$.

Boolean functions for us mean functions whose range is $\{-1,+1\}$.

FOURIER SERIES. We define the inner product of $f, g : \{0,1\}^n \to \mathbf{R}$ by

$$\langle f, g \rangle = 2^{-n} \sum_{x \in \{0,1\}^n} f(x) g(x) .$$

Note that $\langle f, g \rangle = \mathbf{E}[fg]$ where the expectation is over the uniform distribution on the inputs.

The norm associated to this inner product is $\| f \| = \sqrt{\langle f, f \rangle}$.

The *parity functions* are the functions $\chi_z : \{0,1\}^n \to \{-1,+1\}$ defined for $z \in \{0,1\}^n$ by

$$\chi_z(x) = \prod_{i \in z} (-1)^{x_i} .$$

$\{\chi_z\}_{z\in\{0,1\}^n}$ is an orthonormal basis for the vector space of real valued functions on $\{0,1\}^n$ (the orthonormality is with respect to the inner product defined above).

Suppose $f : \{0,1\}^n \to \mathbf{R}$. Then $f = \sum_{z\in\{0,1\}^n} \widehat{f}(z)\chi_z$ where $\widehat{f}(z) = \langle f, \chi_z\rangle$. This (unique) expansion of f in terms of the parity basis is its *Fourier series*. The Fourier coefficients of f are the real numbers $\widehat{f}(z)$ ($z \in \{0,1\}^n$). The sequence of Fourier coefficients is also called the *spectrum* of f. The *Fourier transform* is the operator $\mathcal{F}$ defined by $\mathcal{F}(f) = \widehat{f}$. This operator is linear: $\mathcal{F}(af + g) = a\mathcal{F}(f) + \mathcal{F}(g)$ for real a and functions $f, g : \{0,1\}^n \to \mathbf{R}$.

The *spectral norm* (or just *norm*) of f is $L(f) = \sum_{z\in\{0,1\}^n} |\widehat{f}(z)|$. In the context of norms, "small" means polynomially bounded (as a function of n) and "large" means not small.

Parseval's identity

$$\sum_{z\in\{0,1\}^n} \widehat{f}(z)^2 = \|f\|^2$$

follows directly from the orthonormality of the basis. It follows that if f is boolean then $\sum_{z\in\{0,1\}^n} \widehat{f}(z)^2 = 1$. A consequence of this last fact is the following

Proposition 2.1 *If f is boolean then $L(f) \leq 2^{n/2}$.*

A useful property of the parity functions is that $\chi_a\chi_b = \chi_{a\,\triangle\, b}$.

SOME INTUITION FOR THE BOOLEAN CASE. Some intuition about what the Fourier coefficients of a boolean function f measure is provided by Linial, Mansour and Nisan [LMN]. They observe that $\widehat{f}(z) = \Pr[f(x) = \chi_z(x)] - \Pr[f(x) \neq \chi_z(x)]$ (the probabilities being over a random choice of the input x). Thus one can think of the Fourier coefficient $\widehat{f}(z)$ as measuring the ability of f to approximate the parity of the bits in z when the input is chosen at random.

MATRIX FORMULATION. The following formulation of the Fourier transform in terms of matrices is often useful.

The *Sylvester type Hadamard Matrices* are defined inductively by $H_0 = [1]$ and

$$H_{n+1} = \begin{bmatrix} H_n & H_n \\ H_n & -H_n \end{bmatrix}.$$

Let $[f]$ denote the vector (of length 2^n) whose components are the values of $f(x)$ in lexicographic order. Then

Proposition 2.2 [Br] *Let $f : \{0,1\}^n \to \mathbf{R}$. Then $[\widehat{f}] = 2^{-n}H_n[f]$.*

3 Two Motivating Results

Fourier analysis has of late proved to be a very useful tool in learning and computational complexity theory. And in many recent results, the spectral norm plays a pivotal role. Here are two examples.

3.1 Efficient Learnability and the Spectral Norm

Learning is hard. How hard? The spectral norm of the target function, say Kushilevitz and Mansour [KM], gives us some indication. Let us see how.

THE MODEL. The learning algorithm $\mathcal{A}$ receives as input $1^n, \epsilon, \delta$ and, as an oracle, the function $f : \{0,1\}^n \to \{-1,+1\}$ which it is trying to learn (the latter means that as part of its normal

$\mathcal{A}_B(1^n, \epsilon, \delta; f)$
  $S \leftarrow \mathbf{Coef}(1^n, \lambda, \epsilon B(n)^{-1}; f)$
  $h(\cdot) \leftarrow sign(\sum_{z \in S} \widehat{f}(z)\chi_z)$
  `return` h

$\mathbf{Coef}(1^n, \alpha, \Theta; f)$
{ Returns a superset of $\{ \alpha\beta : |\widehat{f}(\alpha\beta)| \geq \Theta \}$ }
  `if` $\alpha \in \{0,1\}^n$ `then return` $\{\alpha\}$
  `else if` $\mathbf{E}[f_\alpha^2] \geq \Theta^2$ `then return` $\mathbf{Coef}(1^n, \alpha 0, \Theta; f) \cup \mathbf{Coef}(1^n, \alpha 1, \Theta; f)$
    `else return` $\emptyset$

Figure 1: The Learning Algorithm $\mathcal{A}_B$

operation the algorithm can query the value of f on an input x and receive the answer in one step). The output of $\mathcal{A}$ on these inputs is (an encoding of) a *hypothesis* $h : \{0,1\}^n \to \mathbf{R}$. The error of the hypothesis with respect to the input f is $Err_h(f) = \Pr[f(x) \neq h(x)]$ (the probability is over a random choice of x). We can now say what it means to learn a concept class.

Definition 3.1 For each n let $\mathcal{C}^n$ be a collection of functions from $\{0,1\}^n \to \{-1,+1\}$. We say that the concept class $\mathcal{C} = \bigcup_{n \geq 1} \mathcal{C}^n$ is *learnable* if there exists an algorithm $\mathcal{A}$ (of the kind described above) such that $\Pr[Err_{\mathcal{A}(1^n,\epsilon,\delta;f)}(f) \leq \epsilon] \geq 1 - \delta$ for every $f \in \mathcal{C}^n$ and every n and $\epsilon, \delta > 0$ (the probability is over the coin tosses of the algorithm).

Note that this is not distribution free learning in the sense of Valiant [Va]. Not only is the algorithm allowed membership queries but also the error of the hypothesis is measured with respect to the uniform distribution on the inputs.

THE RESULT. Let $\mathcal{C}_B = \bigcup_{n \geq 1} \mathcal{C}_B^n$ where $\mathcal{C}_B^n$ is the class of functions from $\{0,1\}^n \to \{-1,+1\}$ whose spectral norm is bounded above by $B(n)$.

Theorem 3.2 [KM] *Let $B : \mathbf{N} \to \mathbf{N}$. Then the concept class $\mathcal{C}_B$ is learnable in time polynomial in $B(n), n, \epsilon^{-1}$ and $\log \delta^{-1}$.*

Broadly speaking, this says that a concept class is learnable in time polynomial in a bound on the spectral norm of the functions in the class.

THE ALGORITHM. The learning algorithm is presented in Figure 1. Let me explain, very briefly, what is happening.

The idea is to approximate f by the sum of a small number of "special" terms of its Fourier series. These special terms are those which have a high Fourier coefficient: specifically, one can show that if $S \subseteq \{0,1\}^n$ includes all strings z satisfying $|\widehat{f}(z)| \geq \epsilon B(n)^{-1}$ then $\sum_{z \in S} \widehat{f}(z)\chi_z$ is a good approximation to f.

Call z *special* if $|\widehat{f}(z)| \geq \Theta$. The main component of the algorithm is a subroutine **Coef** which finds all the special terms of f: $\mathbf{Coef}(1^n, \lambda, \Theta; f)$ returns a set of n bit strings containing all the special ones (λ is the empty string).

Coef works recursively, building up the special strings from the empty string: if $\alpha \in \{0,1\}^k$ then $\mathbf{Coef}(1^n, \alpha, \Theta; f)$ returns (a superset of) the set of all special strings which have prefix α. The routine

looks at the function $f_\alpha : \{0,1\}^{n-k} \to \{-1,+1\}$ defined by $f_\alpha(x) = \sum_{\beta \in \{0,1\}^{n-k}} \widehat{f}(\alpha\beta)\chi_\beta(x)$: the value of $\mathbf{E}[f_\alpha^2]$ is used to prune the search and keep the running time polynomial.

Finally, let me remark that the algorithm as stated is somewhat inaccurate in that it assumes we can compute both $\mathbf{E}[f_\alpha^2]$ and $\widehat{f}(z)$. The truth, however, is that both can be sufficiently well approximated for our purposes.

PROOFS? They await §7.2 where I will show something more general. The correctness of the above will follow as a special case.

3.2 Sparse Representations and the Spectral Norm

One of the drawbacks of the Fourier series, from a computational point of view, is that it is just too large: the series, after all, has an exponential number of terms. So we seek to approximate functions by sparser Fourier series. But how sparse a series can one get? The answer, again, is in the spectral norm. The result of Siu and Bruck [SB] that I will describe here shows the existence of approximations to a function f by a Fourier series having a number of terms which is polynomial in the spectral norm of f.

Applications abound, particularly in the area of threshold circuit complexity [Br], [HMPST], [BS], [SB].

THE RESULT. The theorem of Siu and Bruck [SB] improves on work in [BS]. The proof I will present here is a slightly modified version of the one in [SB]: I make the probabilistic method more explicit and use Chernoff bounds to avoid the variance calculation.

Lemma 3.3 (Chernoff Bound) *Let $X_1, \ldots, X_m$ be independent, mean 0 random variables in the range $[-1,1]$ and let $X = \sum_{i=1}^m X_i$. Then*

$$\Pr[|X| > A] \;<\; 2e^{-A^2/2m}$$

for any $A > 0$.

Note: Siu and Bruck [SB] assume the domain of their functions is $\{-1,+1\}^n$, rather than $\{0,1\}^n$ as I assume here. So the statement of the following theorem will look a little different from theirs. It is however easy to see that the two formulations are equivalent.

Theorem 3.4 [SB] *Let $f\{0,1\}^n \to \{-1,+1\}$ and let $k > 0$. Then there exists a set $S \subseteq \{0,1\}^n$ and integers M, w_z $(z \in S)$ such that $|f(x) - g(x)| \leq n^{-k}$ where $g : \{0,1\}^n \to \mathbf{R}$ is defined by $g(x) = \frac{1}{M}\sum_{z \in S} w_z \chi_z(x)$. Moreover $|S|, |M|$ are $\leq O(n^{2k+1}L(f)^2)$ and $w_z \in \{-1,+1\}$ $(z \in S)$.*

Proof: The proof uses the probabilistic method. Define $\Omega_i = \{0,1\}^n \cup \{\perp_i\}$ with

$$\Pr_{\Omega_i}(z) \;=\; \begin{cases} |a_z|/B & \text{if } z \in \{0,1\}^n \\ 1 - \sum_{\alpha \in \{0,1\}^n} |a_\alpha|/B & \text{otherwise}, \end{cases}$$

where $B = \lceil L(f) \rceil$ and $a_z = \widehat{f}(z)$. Our sample space will be $\Omega = \prod_{i=1}^N \Omega_i$ with the induced probability measure. Now define random variables

$$Z_{i,x}(z^1, \ldots, z^N) \;=\; \begin{cases} sign(a_{z^i})\chi_{z^i}(x) & \text{if } z^i \in \{0,1\}^n \\ 0 & \text{otherwise} \end{cases}$$

and let $G_x = \frac{B}{N}\sum_{i=1}^{N} Z_{i,x}$. Then

$$\Pr\left[|G_x - f(x)| > n^{-k}\right] = \Pr\left[\left|\sum_{i=1}^{N} Z_{i,x} - \frac{Nf(x)}{B}\right| > \frac{Nn^{-k}}{B}\right],$$

and noting that $\mathbf{E}[Z_{i,x}] = f(x)/B$ this is $< 2e^{-Nn^{-2k}/8B^2}$ by Lemma 3.3. Choosing $N = 16B^2n^{2k+1}$ this is $< 2^{-n}$ and thus

$$\Pr\left[|G_x - f(x)| \leq n^{-k} \text{ for all } x \in \{0,1\}^n\right] > 0 .$$

Thus there is a point $(z^1, \ldots, z^N)$ in our sample space such that $|G_x(z^1, \ldots, z^n) - f(x)| \leq n^{-k}$ for all $x \in \{0,1\}^n$ and we can conclude the proof by setting $g = \frac{B}{N}\sum_{i=1}^{N} sign(a_{z^i})\chi_{z^i}$ (actually the sum is only over the i for which z^i is in $\{0,1\}^n$). ■

The reader is referred to [SB] for applications to the construction of small depth threshold circuits.

4 Upper Bounds: Functions of Small Norm

Convinced that functions of small norm are nice functions, the first thing we would like to do is certainly to find examples of them. In this section I present two simple examples.

AND AND OR. For any $z \in \{0,1\}^n$ we define $\mathrm{AND}_z, \mathrm{OR}_z : \{0,1\}^n \to \{-1,+1\}$ by

$$\begin{aligned} \mathrm{AND}_z(x) &= (-1)^{\bigwedge_{i \in z} x_i} \\ \mathrm{OR}_z(x) &= (-1)^{\bigvee_{i \in z} x_i} . \end{aligned}$$

Proposition 4.1 $L(\mathrm{AND}_z), L(\mathrm{OR}_z) \leq 3$.

Proof: It is not too hard to compute explicitly the complete spectrum of these functions. Details omitted. ■

COMPARISON. Sometimes it is possible to show that a function has small norm by exploiting some intrinsic property of it (for example, a convenient inductive structure). The comparison function is an example [SB].

The comparison function $C_n : \{0,1\}^n \to \{-1,+1\}$ is defined for even n by

$$C_n(xy) = \begin{cases} 1 & \text{if } x \geq y \\ -1 & \text{otherwise} \end{cases}$$

where $x, y \in \{0,1\}^{n/2}$ are identified with the integers $\sum_{i=1}^{n/2} x_i 2^{\frac{n}{2}-i}$ and $\sum_{i=1}^{n/2} y_i 2^{\frac{n}{2}-i}$ respectively.

Proposition 4.2 [SB] *Let n be even. Then $L(C_n) = \frac{n}{2} + 1$.*

Proof: The proof is by induction on $k = n/2$. For the base case $k = 1$ it is easily verified that $C_2 = \frac{1}{2}\chi_{00} + \frac{1}{2}\chi_{01} - \frac{1}{2}\chi_{10} + \frac{1}{2}\chi_{11}$ and hence $L(C_2) = 2$. For $k \geq 1$ we note that

$$C_{2k+2}(xy) = \frac{\chi_{0^{k+1}1^{0^{k}}}(xy) - \chi_{1^{0^{k}}0^{k+1}}(xy)}{2} + \frac{1 + \chi_{1^{0^{k}}1^{0^{k}}}(xy)}{2} C_{2k}(x_2 \ldots x_{k+1} y_2 \ldots y_{k+1}) .$$

It follows that $L(C_{2k+2}) = L(C_{2k}) + 1$. ■

5 The Spectral Norm of Decision Trees

A particularly interesting aspect of spectral norms is how they can relate to the computational complexity of the function in some model of computation. The first (to my knowledge) example is the result of [KM] on parity decision trees.

Here I present a general framework for upper bounding the spectral norm of an arbitrary decision tree in terms of the spectral norms of the functions in its nodes. Extending the result of [KM], this yields many examples of classes of decision trees of small norm.

5.1 A General Framework

THE MODEL. A *decision tree* is a binary tree in which each node is labeled by a boolean function $\{0,1\}^n \to \{-1,+1\}$. The tree computes (or defines) a function $f : \{0,1\}^n \to \{-1,+1\}$ in the natural way. On any input $x \in \{0,1\}^n$, we start at the root and compute the value at x of the function g by which the root is labeled. We branch according to this value: to the left if it was -1 and to the right if it was 1. We continue in this way until we hit a leaf. Here we compute the value at x of the function g by which this leaf is labeled, and this value is the value of the function f on input x.

Note that we do not for now restrict the functions labeling the nodes: *any* boolean function is allowed.

I will identify the decision tree with the function it computes. Thus I will speak of the Fourier transform or the spectral norm of a decision tree, and write things like $\widehat{T}(z)$ or $L(T)$ correspondingly. Similarly I identify a node with the function which labels it and speak of the Fourier transform or norm of a node.

Finally, $|T|$ will denote the number of nodes in T.

DEGREES. We assign to each function a number which we will call its *degree*. The degrees of the constituent nodes will enable us to measure the spectral norm of a decision tree. The definition itself is very simple: we let $deg(g) = \frac{1}{2}[L(g) + 1]$. So the degrees are, strictly speaking, the spectral norms themselves: it is just more convenient to translate slightly.

Note that $L(g) \geq 1$ for any boolean g by Proposition 6.1. So 1 is also the minimal degree of any boolean function.

WEIGHT OF A DECISION TREE. We will assign to each decision tree a weight which is computed as a function of the spectral norms (or, more precisely, the degrees) of its constituent nodes. The weight $w(T)$ of a decision tree T is defined by induction on the structure of the binary tree T as follows:

- If $|T| = 1$ then we let $w(T) = L(g)$ where g is the function labeling the (single) node which comprises the tree.
- If $|T| \geq 3$ then T is as depicted in Figure 2 and we let $w(T) = deg(g)[w(T_{-1}) + w(T_1)]$.

USEFUL FACTS. Before we can prove the main theorem we need to develop some facts about the spectral norm.

Lemma 5.1 *$L(f\chi_s) = L(f)$ for any function f.*

Proof: We have

$$\widehat{f\chi_s}(z) = 2^{-n}\sum_x f(x)\chi_s(x)\chi_z(x) = 2^{-n}\sum_x f(x)\chi_{s\,\triangle\, z}(x) = \widehat{f}(s \triangle z)$$

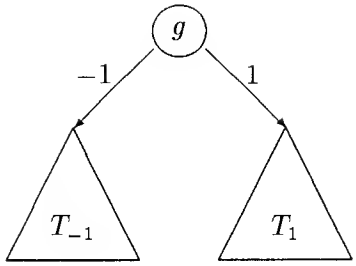

Figure 2: The decision tree T

and the lemma follows from the fact that the map $z \mapsto s \triangle z$ is a permutation of $\{0,1\}^n$. ■

Lemma 5.2 *Suppose* $f, g : \{0,1\}^n \to \mathbf{R}$. *Then*

(1) $L(f+g) \leq L(f) + L(g)$

(2) $L(fg) \leq L(f)L(g)$.

Proof: By the linearity of the fourier transform we have $\mathcal{F}(f+g)(z) = \mathcal{F}(f)(z) + \mathcal{F}(g)(z)$. Apply the triangle inequality and then sum over all z to get (1).

Let $a_z = \widehat{g}(z)$. Then making use of (1) we have

$$L(fg) \;=\; L(f \cdot \textstyle\sum_z a_z \chi_z) \;=\; L(\sum_z a_z f \chi_z) \;\leq\; \sum_z L(a_z f \chi_z) \;=\; \sum_z |a_z| L(f\chi_z)\,.$$

But by Lemma 5.1 this equals

$$\textstyle\sum_z |a_z| L(f) \;=\; L(f) \sum_z |a_z| \;=\; L(f)L(g)$$

which establishes (2). ■

MAIN THEOREM. The main theorem is now very simply stated. It says that the spectral norm of a decision tree is at most its weight.

Theorem 5.3 *Let* T *be a decision tree. Then* $L(T) \leq w(T)$.

Proof: The proof is by induction on the structure of the tree. The base case of T having only one node is easily verified. Now suppose T has ≥ 3 nodes. Visualize T as depicted in Figure 2. Now observe that

$$\begin{aligned} T(x) &= \tfrac{1}{2}[1 - g(x)]T_{-1}(x) + \tfrac{1}{2}[1 + g(x)]T_1(x) \\ &= \tfrac{1}{2}T_{-1}(x) + \tfrac{1}{2}T_1(x) - \tfrac{1}{2}T_{-1}(x)g(x) + \tfrac{1}{2}T_1(x)g(x)\,. \end{aligned}$$

By Lemma 5.2 we get

$$\begin{aligned} L(T) &\leq \tfrac{1}{2}L(T_{-1}) + \tfrac{1}{2}L(T_1) + \tfrac{1}{2}L(T_{-1})L(g) + \tfrac{1}{2}L(T_1)L(g) \\ &= \tfrac{1}{2}\left[L(g)+1\right]L(T_{-1}) \;+\; \tfrac{1}{2}\left[L(g)+1\right]L(T_1) \\ &= deg(g)\left[L(T_{-1}) + L(T_1)\right]\,. \end{aligned}$$

But by induction $L(T_{-1}) \leq w(T_{-1})$ and $L(T_1) \leq w(T_1)$ so the above is $\leq deg(g)[w(T_{-1}) + w(T_1)] = w(T)$. ■

5.2 Some Applications: Decision Trees of Small Norm

Usually, we are interested in decision trees all of whose nodes are from some small *basis* of functions. For example, [KM] consider decision trees where the nodes all compute parity functions. In many of these special cases, the weight of the tree is easy to compute. Theorem 5.3 can then be applied to obtain a useful upper bounds on the spectral norm of the tree.

This technique is quite general, and one can use it to bound the spectral norm of trees with all kinds of combinations of functions in the nodes. Here I illustrate with just two examples.

PARITY DECISION TREES. Let us begin by deriving the result of [KM] that the spectral norm of a *parity* decision tree is bounded above by the number of nodes in the tree (a parity decision tree is a decision tree each of whose internal nodes is labeled by a parity function χ_s and each of whose leaves is labeled $+1$ or -1).

Corollary 5.4 [KM] *Let T be a parity decision tree. Then $L(T) \leq \lceil |T|/2 \rceil$.*

Proof: By the main theorem it suffices to show that the weight of a parity decision tree T is $\leq \lceil |T|/2 \rceil$. The important observation for this is that $deg(\chi_s) = 1$. The fact that $w(T) \leq \lceil |T|/2 \rceil$ is now easily established by induction using the definition of w. ■

CHARACTERIZATION OF DEGREE 1 FUNCTIONS. We note that the only fact about the parity functions that we used in the above is that $deg(\chi_s) = 1$. Thus the theorem extends to decision trees each of whose nodes is labeled by a function of degree 1. As the following characterization of degree 1 functions shows, however, these decision trees are actually identical to parity decision trees.

Proposition 5.5 *Suppose g is boolean. Then $deg(g) = 1$ if and only if $L(g) = 1$ if and only if $g = \pm\chi_s$ for some s.*

Proof: It is clear that $deg(g) = 1$ if and only if $L(g) = 1$, and it is also clear that $L(g) = 1$ if $g = \pm\chi_s$. To complete the proof it suffices to show that if $L(g) = 1$ then $g = \pm\chi_s$ for some s. So suppose $L(g) = 1$. Then we have

$$1 = |g(x)| = |\sum_z \widehat{g}(z)\chi_z(x)| \leq \sum_z |\widehat{g}(z)\chi_z(x)| = L(g) = 1$$

and thus $|\sum_z \widehat{g}(z)\chi_z(x)| = \sum_z |\widehat{g}(z)\chi_z(x)|$. It follows that for each x either $\widehat{g}(z)\chi_z(x) \geq 0$ for all z or $\widehat{g}(z)\chi_z(x) \leq 0$ for all z. Moreover the former happens when $g(x) = 1$ and the latter when $g(x) = -1$, because $g(x) = \sum_z \widehat{g}(z)\chi_z(x)$. So fixing any z such that $\widehat{g}(z) \neq 0$ (at least one such must exist) we can conclude that $sign(\widehat{g}(z)) = \chi_z(x)$ if $g(x) = 1$ and $sign(\widehat{g}(z)) = -\chi_z(x)$ if $g(x) = -1$. But this is true for all x, and thus g is either χ_z or $-\chi_z$. ■

So the parity functions and their negations are exactly the boolean functions of minimal degree.

COMPUTING THE WEIGHT. Before proceeding to more applications let us derive a more useful formulation of the weight function. Namely,

$$w(T) = \sum_{leaves\ l} \prod_{ancestors\ i\ of\ l} deg(i)\,.$$

That is, the weight is the sum, over all leaf to root paths, of the product of the degrees along these paths (we are identifying a node with the function labeling it, so that $deg(i)$ means the degree of the function labeling node i).

AND AND OR. We looked at the spectral norm of the AND and OR functions in §4, and by Proposition 4.1 we know that $deg(\mathrm{AND}_z), deg(\mathrm{OR}_z) \leq 2$. So these functions are just a step higher than parity in the degree hierarchy, and the natural choice of an addition to the basis to get a richer class of decision trees.

Corollary 5.6 *Let T be a decision tree over the basis $\{\chi_s, \mathrm{AND}_s, \mathrm{OR}_s : s \in \{0,1\}^n\}$ with the property that any root to leaf path has at most $O(\log n)$ AND_s or OR_s nodes. Then $L(T) \leq |T| n^{O(1)}$.*

Proof: The weight is easily bounded using the above expression. ■

I don't have space for more examples, but these should have sufficed to illustrate quite well how the main theorem can be applied.

6 Lower Bounds: Functions of Large Norm

I now turn to lower bounds on the spectral norm. I'll first (briefly) consider arbitrary real valued functions and then move on to the real case of interest: boolean functions.

6.1 General Functions

Let's begin with the following observation.

Proposition 6.1 *$L(f) \geq \max_x |f(x)|$ for any function $f : \{0,1\}^n \to \mathbf{R}$.*

Proof: For any x it is the case that

$$L(f) = \sum_z |\widehat{f}(z)| = \sum_z |\widehat{f}(z)\chi_z(x)| \geq |\sum_z \widehat{f}(z)\chi_z(x)| = |f(x)| . \blacksquare$$

This means that for a function to have polynomially bounded spectral norm, its output must be at most $O(\log n)$ bits long. Many simple and interesting functions are thus ruled out. Addition of two n bit numbers is one such.

One quickly realizes however, that the above is not really a restriction. In most applications (learning is an example) the output can be viewed bit by bit. That is, we view the function as a concatenation of boolean functions, and consider the spectral norm of each of these boolean functions separately. So we should concentrate on boolean functions.

6.2 Boolean Functions

It is probably not surprising that

Theorem 6.2 *Most boolean functions have large spectral norm.*

We can view this as a corollary of the fact that functions of small norm are efficiently learnable (Theorem 3.2). A direct proof is not as obvious as it might seem (and I haven't seen one).

Can we construct a *specific* example of a boolean function of large norm? The answer is yes. The function is the "inner product mode 2" (it is "mode," not "mod") defined for even n by

$$I_n(x) = \prod_{i=1}^{n/2} (-1)^{x_{2i-1} \wedge x_{2i}}$$

Theorem 6.3 [MS],[Br] *Let n be even. Then $L(I_n) = 2^{n/2}$.*

Proof: This proof uses the matrix representations of Fourier transforms that we introduced in §2.

We claim that $[I_n]$ is an eigenvector of H_n with eigenvalue $2^{n/2}$. Proposition 2.2 then implies that $[\widehat{I_n}] = 2^{-n/2}[I_n]$, and the result follows. The fact that $[I_n]$ is an eigenvector of H_n with eigenvalue $2^{n/2}$ is easily established by induction on $k = n/2$. ■

This proof is from [Br] who however cites [MS] for the result.

We note that by Proposition 2.1 the bound of Theorem 6.3 is tight.

A few other simple examples of functions of exponential spectral norm are known [Br],[BS]. The proofs in all cases are based on inductive ideas like the ones above (and indeed the functions are constructed to have enough inductive structure to make these proofs go through). I'll now consider a somewhat different kind of boolean function: majority.

6.3 An Exponential Lower Bound for Majority

The majority function $M_n : \{0,1\}^n \to \{-1,+1\}$ is defined by

$$M_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n x_i \geq \frac{n}{2} \\ -1 & \text{otherwise.} \end{cases}$$

WHAT'S KNOWN. Siu and Bruck [SB] show that the spectral norm of M_n is not polynomially bounded. Their proof relies on Theorem 3.4 and results from [Br] and [HMPST]. The results from [Br] and [HMPST] are first used to show that M_n is not approximable (to within $o(n^{-1})$ error) by any sparse Fourier series. That $L(M_n)$ is not polynomially bounded then follows from Theorem 3.4.

Assuming this was the best known ([SB] did not indicate otherwise) I proved the exponential lower bound which appears as Theorem 6.4 below. Later I read [BOH] in which, in a different context, appear equations for the *complete spectrum* of the majority function. A lower bound at least as good as Theorem 6.4 is easily derived from this.

They attribute the spectrum to [Ka]. I was however unable to find [Ka] anywhere, and thus do not know the proof of the result stated in [BOH].

I should note, though, that [SB],[BOH] all consider only the case of n being odd (they are interested in majority as a linear threshold function and the case of n even is inconvenient because $\sum_i x_i$ could be 0). My proof applies (only) to n even.

Before going on it is also worth noting that techniques along the lines of those used in the proof of Theorem 6.3 would not appear to be useful here because the majority function has no clear inductive structure.

THE LOWER BOUND. The following exponential lower bound is within a polynomial factor of optimal:

Theorem 6.4 *Let $n \geq 1$ be a multiple of 4. Then*

$$L(M_n) \geq \frac{4e^{-1/n}}{\sqrt{2\pi n}} 2^{n/2} .$$

Let's proceed to the proof.

The attack is pretty direct. What I will do is compute explicitly the "middle" Fourier coefficients (that is, those corresponding to the parity over a size $n/2$ subset of the variables), and get a bound based on these alone.

A COMBINATORIAL IDENTITY. I will need the fact that

$$\sum_{i=0}^{2k} \binom{2k}{i}^2 (-1)^i = (-1)^k \binom{2k}{k} .$$

This is probably known, although I haven't been able to find a reference. For completeness, I provide in Appendix A a simple proof via generating functions.

COMPUTING THE MIDDLE FOURIER COEFFICIENTS. The following lemma gives explicit values for the middle Fourier coefficients of the majority function.

Lemma 6.5 *Let $z \in \{0,1\}^{4k}$ have exactly $2k$ ones. Then*

$$\langle M_{4k}, \chi_z \rangle = 2^{-4k} \binom{2k}{k} (-1)^k .$$

Proof: We begin by observing that the value of the fourier coefficient $\langle M_{4k}, \chi_z \rangle$ depends only on the number of ones in z. So without loss of generality we can set $z = 1^{2k}0^{2k}$. Now let $A = 2^{4k}\langle M_{4k}, \chi_z \rangle$. We claim that

$$A = \sum_{i=0}^{2k} \binom{2k}{i}^2 (-1)^i , \tag{1}$$

whence the result follows by our combinatorial identity.

To establish Equation (1) we begin by noting that $M_{4k}(x)\chi_z(x) = (-1)^i M_{4k}(1^i 0^{2k-i} 1^j 0^{2k-j})$ where i is the number of ones in the left half of x and j is the number of ones in the right half of x. It follows that

$$\begin{aligned} A &= \sum_{i=0}^{2k} \sum_{j=0}^{2k} \binom{2k}{i} \binom{2k}{j} (-1)^i M_{4k}(1^i 0^{2k-i} 1^j 0^{2k-j}) \\ &= \sum_{i=0}^{2k} \binom{2k}{i} (-1)^i \left[\sum_{j=2k-i}^{2k} \binom{2k}{j} - \sum_{j=0}^{2k-i-1} \binom{2k}{j} \right] . \end{aligned}$$

The $i = k$ term of the above sum equals $\binom{2k}{k}^2 (-1)^k$. Let B be the remaining part of the sum. Noting that $\binom{2k}{i}(-1)^i = \binom{2k}{2k-i}(-1)^{2k-i}$ we can rewrite B as

$$\sum_{i=0}^{k-1} \binom{2k}{i} (-1)^i \left[\sum_{j=2k-i}^{2k} \binom{2k}{j} - \sum_{j=0}^{2k-i-1} \binom{2k}{j} + \sum_{j=i}^{2k} \binom{2k}{j} - \sum_{j=0}^{i-1} \binom{2k}{j} \right] .$$

Using the fact that $\binom{2k}{j} = \binom{2k}{2k-j}$ the term in square brackets simplifies to

$$\sum_{j=0}^{i} \binom{2k}{j} - \sum_{j=i+1}^{2k} \binom{2k}{j} + \sum_{j=i}^{2k} \binom{2k}{j} - \sum_{j=0}^{i-1} \binom{2k}{j} = 2\binom{2k}{i} ,$$

and putting all this together we get

$$A = \binom{2k}{k}^2 (-1)^k + 2\sum_{i=0}^{k-1} \binom{2k}{i}^2 (-1)^i = \sum_{i=0}^{2k} \binom{2k}{i}^2 (-1)^i$$

as desired. ■

PROOF OF THEOREM 6.4. The proof of Theorem 6.4 follows easily now. First let us recall the bound

$$\binom{2n}{n} \geq \frac{e^{-1/6n}}{\sqrt{\pi n}} 2^{2n} \tag{2}$$

(which is obtained using Stirling's formula). Now let $k = n/4$. Then $L(M_{4k}) = \sum_z |\langle M_{4k}, \chi_z \rangle| \geq \sum_{|z|=2k} |\langle M_{4k}, \chi_z \rangle|$, and this by Lemma 6.5 and Equation (2) is at least

$$\binom{4k}{2k} 2^{-4k} \binom{2k}{k} \geq \frac{e^{-1/4k}}{\sqrt{2}\pi k} 2^{2k} = \frac{4e^{-1/n}}{\sqrt{2}\pi n} 2^{n/2}$$

as desired.

7 Generalization: q-norms

I now turn to a generalization of the basic Fourier series and the corresponding spectral norms (q-norms). Results about the q-norm being much scarcer in the literature (it is a recent invention of Furst, Jackson and Smith [FJS]) I will spend more time here on developing techniques and applications. In particular, I present a new application.

The motivation, once again, comes from learning.

One of the drawbacks of the [KM] learning algorithm is that the quality of the hypothesis is measured by its proximity to the target under one particular fixed distribution: the uniform one. Can we move any closer to some kind of distribution free learning?

History, luckily, provides a parallel. Linial, Mansour and Nisan [LMN] had used Fourier analysis to develop an algorithm to learn AC^0 functions under the uniform distribution. Furst, Jackson and Smith extended this result to what they called *mutually independent* distributions. In order to do this they developed a more general Fourier transform in which the basis is defined in terms of the mutually independent distribution q.

The tools and framework introduced by [FJS] provide an avenue to similarly generalize the [KM] learning algorithm. Measuring the quality of a hypothesis by its proximity to the target under q, we will be able to learn in time polynomial in the q-norm.

Here I will present this generalization of [KM] and then try to say something about q-norms and their relation to the usual norm.

7.1 The q-Framework

I must begin by developing the machinery. Most of the work, in fact, is here: once things have been properly set up the lemmas of [KM] can be generalized in a pretty straightforward manner.

MUTUALLY INDEPENDENT DISTRIBUTIONS. A probability distribution $q : \{0,1\}^n \to [0,1]$ is *mutually*

independent if the random variables $x_1, \ldots, x_n$ are independent.

We will be assuming that $0 < \Pr[x_i = 1] < 1$ for each i. We say that q is polynomially bounded if

$$\max_{1 \le i \le n} \left[\frac{1}{\Pr[x_i = 1]} + \frac{1}{\Pr[x_i = 0]} \right]$$

is polynomially bounded (as a function of n).

q-FOURIER SERIES. Let q be mutually independent. We define the inner product of $f, g : \{0,1\}^n \to \mathbf{R}$ by

$$\langle f, g \rangle_q = \textstyle\sum_{x \in \{0,1\}^n} f(x) g(x) q(x) .$$

Note that $\langle f, g \rangle_q = \mathbf{E}[fg]$ where the expectation is over q.

The norm associated to this inner product is $\| f \|_q = \sqrt{\langle f, f \rangle_q}$.

Let $\mu_i = \Pr[x_i = 1]$. Note that μ_i is the mean of the random variable x_i. Let the standard deviation $\sqrt{\mu_i(1-\mu_i)}$ of x_i be denoted by σ_i.

For $z, x \in \{0,1\}^n$ we define

$$\phi(z, x) = \prod_{i \in z} \frac{\mu_i - x_i}{\sigma_i} .$$

We call ϕ the basis associated to q because of the following

Proposition 7.1 [Ba],[FJS] *Let $q : \{0,1\}^n \to [0,1]$ be mutually independent and let ϕ be as defined above. Then $\{\phi(z, \cdot)\}_{z \in \{0,1\}^n}$ is an orthonormal basis for the vector space of real valued functions on $\{0,1\}^n$ (the orthonormality is with respect to the inner product $\langle \cdot, \cdot \rangle_q$).*

Note that $\phi(z, \cdot) = \chi_z$ when q is the uniform distribution.

To get a feel for these definitions, let's check the orthonormality. For $a, b \in \{0,1\}^n$ we have

$$\langle \phi(a, \cdot), \phi(b, \cdot) \rangle_q = \mathbf{E}\left[\prod_{i \in a} \frac{\mu_i - x_i}{\sigma_i} \prod_{j \in b} \frac{\mu_j - x_j}{\sigma_j} \right] = \mathbf{E}\left[\prod_{i \in a \cap b} \left(\frac{\mu_i - x_i}{\sigma_i} \right)^2 \prod_{i \in a \,\triangle\, b} \frac{\mu_i - x_i}{\sigma_i} \right] ,$$

and the mutual independence implies that this equals

$$\prod_{i \in a \cap b} \mathbf{E}\left[\left(\frac{\mu_i - x_i}{\sigma_i} \right)^2 \right] \prod_{i \in a \,\triangle\, b} \mathbf{E}\left[\frac{\mu_i - x_i}{\sigma_i} \right] .$$

Now each term of the first product is 1 while each term of the second is 0, and thus the above expression is 1 if $a = b$ and 0 otherwise.

The Fourier series of $f : \{0,1\}^n \to \mathbf{R}$ is $\sum_{z \in \{0,1\}^n} \widehat{f}(z) \phi(z, x)$ where $\widehat{f}(z) = \langle f, \phi(z, \cdot) \rangle_q$. The q-spectral norm (or just q-norm) is $L_q(f) = \sum_{z \in \{0,1\}^n} |\widehat{f}(z)|$. Parseval's identity as usual says that $\sum_{z \in \{0,1\}^n} \widehat{f}(z)^2 = \| f \|_q^2$.

I'm abusing notation a little by denoting by $\widehat{f}(z)$ both the Fourier coefficients under q and the (usual) Fourier coefficients with respect to $\{\chi_z\}$. The context should usually make it clear which I mean, and if both are around I will disambiguate.

THE SHIFTED DISTRIBUTION AND BASIS. With minor changes the above is the framework of [FJS]. I now have to go a little further.

Let q be mutually independent and ϕ the associated basis. Suppose $1 \leq k \leq n$.

For $y \in \{0,1\}^k$ and $0 \leq j \leq n-k$ we define $q_j(y) = \Pr[x_{j+1} = y_1, \ldots, x_{j+k} = y_k]$. Observe that q_j is a mutually independent distribution on $\{0,1\}^k$.

For notational convenience we identify q_0 with q.

For $x, z \in \{0,1\}^k$ and $0 \leq j \leq n-k$ we define

$$\phi_j(z,x) = \prod_{i \in z} \frac{\mu_{j+i} - x_i}{\sigma_{j+i}} .$$

We also define $\phi_j(\lambda, \lambda) = 1$ for all $j = 0, \ldots, n$.

For notational convenience we identify ϕ_0 with ϕ.

The fact that q_j is a mutually independent distribution on $\{0,1\}^k$ implies, by Proposition 7.1, that $\{\phi_j(z,\cdot)\}_{z \in \{0,1\}^k}$ is an orthonormal basis for the vector space of real valued functions on $\{0,1\}^k$ (the orthonormality is with respect to the inner product $\langle \cdot, \cdot \rangle_{q_j}$).

We'll denote by $\mathbf{E}_j$ the expectation over q_j.

A SWITCHING LEMMA. One very useful property of the ϕ-basis is that

$$\phi_j(a,b)\sqrt{q_j(b)} = \phi_j(b,a)\sqrt{q_j(a)}$$

for any $a, b \in \{0,1\}^k$ and any $j = 0, \ldots, n-k$.

7.2 Efficient Learnability and the q-Norm

Fix a mutually independent distribution q. The only change in the learning model is that the probability in the definition of the error $Err_h(f) = \Pr[f(x) \neq h(x)]$ is now taken over q (rather than over x chosen at random).

THE RESULT. Let $\mathcal{C}_B(q) = \bigcup_{n \geq 1} \mathcal{C}_B^n(q)$ where $\mathcal{C}_B^n(q)$ is the class of functions from $\{0,1\}^n \to \{-1,+1\}$ whose spectral q-norm is bounded above by $B(n)$.

Theorem 7.2 *Let $B : \mathbf{N} \to \mathbf{N}$. Let q be a polynomially bounded, mutually independent distribution. Then the concept class $\mathcal{C}_B(q)$ is learnable in time polynomial in $B(n), n, \epsilon^{-1}$ and $\log \delta^{-1}$.*

Note that when q is the uniform distribution this reduces to the [KM] result.

The algorithm is shown in Figure 3. Again, one can check that when q is the uniform distribution it is just the algorithm of Figure 1.

To show that it works, I now generalize the three basic lemmas of [KM]. I'll state and prove the lemmas without elaborating overly; the discussion in §3.1 should suffice to see why these lemmas imply the correctness of the algorithm, and for even more detail the reader is referred to [KM].

THE LEMMAS. Fix a mutually independent distribution q and its associated basis ϕ. Let $f : \{0,1\}^n \to \mathbf{R}$ (note we are not requiring f to be boolean). Suppose $0 \leq k \leq n-1$ and $\alpha \in \{0,1\}^k$. Define $f_\alpha : \{0,1\}^{n-k} \to \mathbf{R}$ by $f_\alpha(x) = \sum_{\beta \in \{0,1\}^{n-k}} \widehat{f}(\alpha\beta)\phi_k(\beta, x)$.

The first lemma shows that it suffices to find the terms of sufficiently large Fourier coefficient. The proof remains identical to the one in [KM].

$\mathcal{A}_{B,q}(1^n, \epsilon, \delta; f)$
  $S \leftarrow \mathbf{Coef}(1^n, \lambda, \epsilon B(n)^{-1}; f)$
  $h(\cdot) \leftarrow sign\,(\sum_{z \in S} \widehat{f}(z)\phi(z, \cdot))$
  `return` h

$\mathbf{Coef}(1^n, \alpha, \Theta; f)$
{ Returns a superset of $\{ \alpha\beta : |\widehat{f}(\alpha\beta)| \geq \Theta \}$ }
  `if` $\alpha \in \{0,1\}^n$ `then return` $\{\alpha\}$
  `else` Let k be the length of α
    `if` $\mathbf{E}_k[f_\alpha^2] \geq \Theta^2$ `then return` $\mathbf{Coef}(1^n, \alpha 0, \Theta; f) \cup \mathbf{Coef}(1^n, \alpha 1, \Theta; f)$
    `else return` $\emptyset$

Figure 3: The Generalized Learning Algorithm $\mathcal{A}_{B,q}$

Lemma 7.3 *Let ϵ be > 0 and suppose $S \supseteq \{ z : |\widehat{f}(z)| \geq \epsilon L_q(f)^{-1} \}$. Let $g(x) = \sum_{x \in S} \widehat{f}(z)\phi(z,x)$. Then $\mathbf{E}[(f-g)^2] \leq \epsilon$. Moreover if f is boolean then $\Pr[f(x) \neq \text{sign}\,(g(x))] \leq \epsilon$.*

Proof:

$$\mathbf{E}[(f-g)^2] \;=\; \sum_{z \in \{0,1\}^n} [\widehat{f}(z) - \widehat{g}(z)]^2 \;=\; \sum_{z \notin S} \widehat{f}(z)^2 \;,$$

and the last of these is at most

$$\max_{z \notin S} |\widehat{f}(z)| \cdot \sum_{z \in \{0,1\}^n} |\widehat{f}(z)| \;\leq\; \epsilon L_q(f)^{-1} \cdot L_q(f) \;=\; \epsilon \;.$$

For boolean f we note that $\Pr[f(x) \neq sign(g(x))] \leq \Pr[|f(x) - g(x)| > 1] \leq \mathbf{E}[(f-g)^2]$. ■

The next lemma is used to bound the search time. I state it slightly more generally than [KM] so that the case of non-boolean functions is covered.

Lemma 7.4 *At most a $\| f \|_q^2 \, \Theta^{-2}$ fraction of the funetions f_α ($\alpha \in \{0,1\}^k$) have $\mathbf{E}_k[f_\alpha^2] \geq \Theta^2$.*

Proof: We observe that

$$\sum_\alpha \mathbf{E}_k[f_\alpha^2] \;=\; \sum_\alpha \langle f_\alpha, f_\alpha \rangle_{q_k} \;=\; \sum_\alpha \left\langle \textstyle\sum_\beta \widehat{f}(\alpha\beta)\phi_k(\beta, \cdot)\,,\, \sum_\beta \widehat{f}(\alpha\beta)\phi_k(\beta, \cdot) \right\rangle_{q_k}$$

which by the bilinearity of the inner product and the orthonormality of the basis is

$$\sum_{\alpha\beta} \widehat{f}(\alpha\beta)^2 \;= \| f \|_q^2 \;.$$

The lemma follows. ■

Note that $\| f \|_q \leq L_q(f)$ and hence if $L_q(f)$ is polynomially bounded then so is $\| f \|_q$.

Finally we have to show how to approximate $\mathbf{E}_k[f_\alpha]$. It can be done by sampling provided we can approximate f_α itself. By equating f_α with an expectation the following lemma provides the avenue to obtain the necessary approximation to it.

Lemma 7.5 $f_\alpha(x) = \mathbf{E}[f(\cdot\, x)\phi(\alpha, \cdot)]$.

Proof: Making liberal use of the identities in §7.1 we have

$$\begin{aligned}
f_\alpha(x) &= \sum_\beta \widehat{f}(\alpha\beta)\phi_k(\beta,x) \\
&= \sum_\beta \sum_{yz} f(yz)\phi(\alpha\beta,yz)q(yz)\phi_k(\beta,x) \\
&= \sum_{yz} f(yz)\phi_0(\alpha,y)q(yz)\sum_\beta \phi_k(\beta,z)\phi_k(\beta,x) \\
&= \sum_{yz} f(yz)\phi_0(\alpha,y)q(yz)\sum_\beta \phi_k(z,\beta)\sqrt{\frac{q_k(\beta)}{q_k(z)}}\,\phi_k(x,\beta)\sqrt{\frac{q_k(\beta)}{q_k(x)}} \\
&= \sum_{yz} f(yz)\phi_0(\alpha,y)q_0(y)\sqrt{\frac{q_k(z)}{q_k(x)}}\,\langle\phi_k(z,\cdot),\phi_k(x,\cdot)\rangle_{q_k} \\
&= \sum_y f(yx)\phi_0(\alpha,y)q_0(y) \\
&= \mathbf{E}[f(\cdot\, x)\phi(\alpha,\cdot)]\ . \quad \blacksquare
\end{aligned}$$

LEARNING NON-BOOLEAN FUNCTIONS. The restriction that the target function be boolean is not really necessary. For the general case we change the definition of the error to $Err_h(f) = \mathbf{E}[(f-h)^2]$ (the expectation, as usual, being over q). We also require the target functions to have polynomially bounded output length under some encoding of the output into bits. Finally the algorithm of Figure 3 is modified to output the function $g(\cdot) = \sum_{z\in S} \widehat{f}(z)\phi_z(\cdot)$ rather than just its sign.

7.3 The q-norm of Decision Trees

The results of §5 can also be generalized by appropriately extending the definition of the degree.

Namely, let q be mutually independent and $g : \{0,1\}^n \to \mathbf{R}$. Then the degree of g with respect to q, denoted $deg_q(g)$, is defined to be the least positive real number d with the property that $L_q(fg) \leq (2d-1)L_q(g)$ for all $f : \{0,1\}^n \to \{-1,+1\}$.

It should be noted that when q is the uniform distribution this definition of the degree coincides with the one in §5. However I suspect that in general it is *not* true that $deg_q(g) = \frac{1}{2}[L_q(g)+1]$.

The weight $w_q(T)$ of a decision tree T is then be defined by induction as before, using the new definition of the degree, and the proof of the main theorem can be extended to show that $L_q(T) \leq w_q(T)$ for any decision tree T. I omit the details.

7.4 L_q versus L

Results such as the above motivate us to understand the relationship between the usual spectral norm and the q-spectral norm. In particular, can the change of basis and distribution drastically reduce the spectral norm?

Proposition 7.6 *There is a function of exponential spectral norm which has polynomial spectral q-norm with respect to some polynomially bounded, mutually independent q.*

Proof: Let q be defined by $\Pr[x_i = 1] = \frac{3}{4}$ for each $i = 1,\ldots,n$. Let ϕ be the associated basis. Then $L(\phi(z,\cdot)) \geq 3^{|z|/2}$ (this follows from Proposition 6.1 since $|\phi(z,0^n)| = 3^{|z|/2}$). So for z having, say, a constant fraction of ones, the (usual) spectral norm of $\phi(z,\cdot)$ is exponential.

The q-norm of $\phi(z,\cdot)$, on the other hand, is of course just 1 since $\phi(z,\cdot)$ is a basis function. ■

So $\phi(z,\cdot)$ is not learnable by the original algorithm but can be learned by the generalized one.

PROBLEM. Find an example of a *boolean* function f and a polynomially bounded, mutually independent q such that $L(f)$ is not polynomially bounded but $L_q(f)$ is.

I don't know how hard this is; I haven't had time to try.

Acknowledgements

Thanks to Ron Rivest for suggesting (and writing) a lisp routine for evaluating $\sum_{i=0}^{2k} \binom{2k}{i}^2 (-1)^i$ — I wouldn't have got the answer without it!

References

[Ba] Bahadur, R,. "A Representation of the Joint Distribution of Respones to n Dichotomous Items," *Studies in Item Analysis and Prediction* (edited by H. Solomon), Stanford University Press (1961).

[BOH] Brandman, Y., A. Orlitsky and John Hennessy, "A Spectral Lower Bound Technique for the Size of Decision Trees and and Two-Level AND/OR Circuits," *IEEE Transactions on Computers* **39**(2), 282-287 (February 1990).

[Br] Bruck, J., "Harmonic Analysis of Polynomial Threshold Functions," *SIAM J. Discrete Math.* **3**(2), 168-177 (1990).

[BS] Bruck, J. and R. Smolensky, "Polynomial Threshold Functions, AC^0 Functions, and Spectral Norms," *Proceedings of the 31st Annual IEEE Symposium on the Foundations of Computer Science,* IEEE (1990).

[FJS] Furst, M., J. Jackson and S. Smith, "Learning AC^0 Functions Sampled under Mutually Independent Distributions," Manuscript (October 1990).

[HMPST] Hajnal, A., W. Maass, P. Pudlak, M. Szegedy and G. Turan, "Threshold Circuits of Bounded Depth," *Proceedings of the 28th Annual IEEE Symposium on the Foundations of Computer Science,* IEEE (1987).

[Ka] Karpovsky, M., "Finite Orthogonal Series in the Design of Digital Devices," Wiley (1976).

[KM] Kushilevitz, E. and Y. Mansour, "Learning Decision Trees using the Fourier Spectrum," Manuscript (November 1990).

[LMN] Linial, N., Y. Mansour and N. Nisan, "Constant Depth Circuits, Fourier Transform, and Learnability," Manuscript (January 1991). Preliminary version in *Proceedings of the 30th Annual IEEE Symposium on the Foundations of Computer Science,* IEEE (1989).

[MS] MacWilliams, F. and N. Sloane, "The Theory of Error-Correcting Codes," North-Holland (1978).

[SB] Siu, K. and J. Bruck, "On the Power of Threshold Circuits with Small Weights," Manuscript.

[Va] Valiant, L., "A Theory of the Learnable," *Communications of the ACM* **27**(11), 1134-1142 (1984).

A Appendix: Proof of the Combinatorial Identity

Here I give a proof of the combinatorial identity used to derive the lower bound on the spectral norm of the majority function. Let's begin by recalling the identity:

$$\sum_{i=0}^{2k} \binom{2k}{i}^2 (-1)^i = (-1)^k \binom{2k}{k} .$$

The proof uses generating functions. We define $a_i = \binom{2k}{i}$ and $b_i = \binom{2k}{i}(-1)^i$ and let

$$G(x) = \left(\sum_{i=0}^{\infty} a_i x^i\right) \cdot \left(\sum_{i=0}^{\infty} b_i x^i\right) = \left(\sum_{i=0}^{\infty} \binom{2k}{i} x^i\right) \cdot \left(\sum_{i=0}^{\infty} \binom{2k}{i} (-1)^i x^i\right) .$$

Let $\sum_{i=0}^{\infty} c_i x^i$ be the formal power series for G. By definition of G we have

$$c_{2k} = \sum_{i=0}^{2k} a_i b_{2k-i} = \sum_{i=0}^{2k} \binom{2k}{i}^2 (-1)^i .$$

So it suffices to show that $c_{2k} = (-1)^k \binom{2k}{k}$. To do this we compute $G(x)$ explicitly. By the binomial theorem $\sum_{i=0}^{\infty} \binom{2k}{i} x^i = (1+x)^{2k}$ and $\sum_{i=0}^{\infty} \binom{2k}{i}(-1)^i x^i = (1-x)^{2k}$, so

$$G(x) = (1+x)^{2k}(1-x)^{2k} = (1-x^2)^{2k} .$$

Applying the binomial theorem again we get

$$G(x) = \sum_{i=0}^{\infty} \binom{2k}{i} (-1)^i x^{2i} ,$$

and thus $c_{2k} = (-1)^k \binom{2k}{k}$ as desired.